AT&T Labs - Research

subject: **Controlling the Impact of BGP Policy Changes on IP Traffic**

date: **November 6, 2001**

from: **Nick Feamster**
MIT
feamster@lcs.mit.edu

Jay Borkenhagen
Dept. HA9215000
MT C5-3D12
732-420-2526
jayb@att.com

Jennifer Rexford
Dept. HA173000
FP A169
973-360-8728
jrex@research.att.com
HA173000-011106-02TM

TECHNICAL MEMORANDUM

The Internet consists of nearly 12,000 autonomous systems (AS's) that exchange routing information using the Border Gateway Protocol (BGP). The operators of each network need to have control over the flow of traffic through the AS. However, BGP does not facilitate common traffic engineering tasks, such as balancing load across multiple links to a neighboring AS or directing traffic to a different neighbor. Solving these problems is difficult because the number of possible changes to routing policies is too large to exhaustively test all possibilities, some changes in routing policy can have an unpredictable effect on the flow of traffic, and the BGP decision process implemented by router vendors limits an operator's control over path selection. In this paper, we demonstrate that it is possible to *predictably* model the changes in traffic flows in response to BGP policy changes, *given that policies are adapted in a certain fashion.* Based on analysis of routing tables and traffic measurements from the AT&T backbone, we show that operators can control the scale of the traffic engineering problem by focusing on the small fraction of destination prefixes (and sets of related prefixes) responsible for the majority of traffic. Furthermore, they can make the effects of their changes more predictable by following specific policy guidelines and selecting configuration options that make the BGP decision process deterministic. This allows an operator to gain more control over network traffic *within the existing BGP framework.*

1 Introduction

Operating a large IP backbone requires continuous attention to the distribution of traffic over the network. Equipment failures and changes in routing policies in neighboring domains can trigger sudden shifts in the flow of traffic. Flash crowds caused by special events and popular new applications can also cause significant changes in the load on the network. Network failures and traffic fluctuations degrade user performance and lead to inefficient use of network resources by leading to unstable network paths and unpredictable round-trip times [1] . Network operators adapt to changes in the distribution of traffic by adjusting the configuration of the routing protocols running on their routers. Additionally, the addition of new routers and links to the network often requires changes in routing configuration. Developing effective techniques for adapting routes to the prevailing traffic and topology has been an active area of research and standards activity during the past few years [2–6] . Previous work has

focused on Interior Gateway Protocols (IGPs), such as OSPF, IS-IS, and MPLS, which control the flow of traffic within a single administrative domain. However, most traffic in a large backbone network traverses multiple domains, making interdomain routing an important part of traffic engineering. In this paper, we address the challenges of using *inter*domain routing policies to control the flow of traffic in an efficient and predictable manner.

The Internet consists of nearly 12,000 autonomous systems (AS's), where an AS is a collection of routers and links administered by an institution, such as a company, university, or Internet service provider (ISP). Neighboring AS's use the Border Gateway Protocol (BGP) to exchange routing information to provide end-to-end connectivity between hosts in different domains [7-9]. Each BGP advertisement announces reachability to a *prefix* that represents a block of IP addresses. Each advertisement includes a list of the AS's in the path to that prefix, as well as a number of other attributes. The routers in each AS apply local *routing policies* that manipulate the attributes associated with these advertisements. In this way, network operators use routing policy to influence the selection of the best route for a particular prefix and to decide whether to propagate this route to neighboring AS's. BGP differs from IGPs that select paths based on link metrics, such as static weights or dynamic load information, because BGP advertisements do not explicitly convey any information about the resource availability on a path. In addition, BGP routing policies are complex and are determined by a variety of factors, such as the commercial relationships with neighboring AS's [10]. Despite the constraints that BGP imposes on making "intelligent" routing decisions, moving to a radically different interdomain routing paradigm would be extremely difficult in practice. Thus, rather than proposing changes or extensions to BGP, we investigate ways to support traffic engineering *within the existing BGP framework*.

Operators influence the flow of traffic across an AS indirectly by tuning the routing policies that affect the selection of the best path for a destination prefix. Choosing the appropriate configuration is difficult since it depends on the network topology (the connectivity between the routers and the capacity of the links, as well as the association of the edge links with particular neighboring AS's), the BGP advertisements from neighboring AS's, and the current traffic patterns. In our work, we focus on the impact of BGP policies on the flow of traffic *leaving* the network at the egress points that connect to neighboring domains. For example, an operator can direct certain traffic to a different next-hop AS by selecting BGP policies that assign a higher preference to advertisements from that AS. Alternatively, an operator may direct traffic to a different egress point to the same next-hop AS to exploit a new link of higher capacity. Some traffic engineering tasks necessitate changes to how traffic *enters* the network. However, we believe that controlling how traffic enters the network in a predictable way requires coordination with neighboring domains. Our techniques for controlling outbound traffic can be applied by the neighboring AS's to influence how traffic enters the network.

Network operators adjust BGP policies and IGP weights to achieve some performance objective, such as balanced link load or bounded propagation delay on each path. Previous work on optimizing intradomain routing has focused on minimizing the utilization of the most heavily-loaded link in the network or minimizing the weighted sum of some function of the load on each link. To capture all of the costs associated with assigning traffic flows to links, an objective function should also incorporate other constraints, such as the need to have a relatively even exchange of traffic to and from particular neighbor AS's. In practice, optimizing the configuration of the IGP parameters is computationally challenging [11]. Allowing changes to BGP policies introduces significantly more complexity to the optimization problem for three reasons. First, router vendors offer a wide array of configuration commands that provide network operators significant flexibility in specifying BGP policies. Second, the selection of the best path for each prefix depends not only on the local routing policies but also on the advertisements sent by neighboring domains. Third, changing the BGP policy in one AS may alter the advertisements propagated to neighboring domains, which may inadvertently affect how traffic enters the AS, making the inbound traffic patterns less predictable.

We propose several ways to scope the BGP traffic engineering problem, based on our analysis of routing and traffic data from a large operational network. Section 2 presents an overview of BGP from the viewpoint of a network operator and describes the steps involved in choosing the best route for each destination prefix. In Section 3, we describe how to decouple the influence of BGP policies and IGP parameters on the path selection process. We propose a simple network-wide representation of the BGP advertisements and describe how to glean this information from BGP routing tables. Additionally, we present a set of principles for ensuring that changes in BGP policy have a *predictable* impact on the flow of traffic without introducing significant instability into the network. Section 4 analyzes routing tables and flow-level traffic measurements from the AT&T IP backbone to identify effective techniques for adapting BGP routing policies to the prevailing traffic. Section 5 presents a summary of the paper and discusses possible avenues for future work on interdomain traffic engineering within the context of BGP.

2 Border Gateway Protocol

In this section, we present an overview of BGP and the attributes associated with BGP advertisements. We then describe how a router selects a best path for each block of IP addresses when constructing a forwarding table.

2.1 BGP Protocol

Internet routing operates at the level of address blocks, or prefixes. Each prefix consists of a 32-bit address and a mask length; for example, 192.0.2.0/24 consists of 256 addresses ranging from 192.0.2.0 to 192.0.2.255. An IP router constructs a forwarding table that is used to select the output interface for each incoming packet, based on the longest-matching prefix for that destination address. Routers in different AS's use BGP to exchange update messages about how to reach different destination prefixes. A router sends an *announcement* to notify its neighbor of a new route to the destination prefix and sends a *withdrawal* to revoke the route when it is no longer available. Each advertisement includes a number of attributes about the route, including the list of AS's along the path to the destination prefix. Before accepting an advertisement, the receiving router checks for the presence of its own AS number in the AS path to detect and remove routing loops.

A router may receive routes for the destination prefix from multiple neighboring AS's. The router applies *import policies* to filter unwanted routes and to manipulate the attributes of the remaining routes. Ultimately, the router invokes a *decision process* to select exactly one "best" route for each destination prefix among all the routes it hears. The router then applies *export policies* to manipulate attributes and decide whether to advertise the route to neighboring AS's. Router vendors provide a large number of configuration commands for composing the import and export policies. In addition to exchanging BGP messages with neighboring domains, an AS may use internal BGP (iBGP) to distribute routing information amongst its routers[1]. Ultimately, every router must select a single best route for each prefix among the advertisements from the various eBGP (external BGP) and iBGP neighbors.

BGP advertisements can include numerous attributes [8], including:

- *AS path*: The AS path identifies the list of AS's en route to the origin AS responsible for the destination prefix.
- *Next hop:* The next hop is the IP address of the border router associated with the path. The IGP dictates how the router would direct traffic toward that egress point.
- *Origin type:* The origin type identifies how the origin AS learned about the route—within the AS (e.g., static configuration), EGP (a now-defunct distance-vector protocol), or injection from another routing protocol. These origin types are known as IGP, EGP, and INCOMPLETE.
- *Multiple exit discriminator:* A BGP advertisement may also include a multiple exit discriminator (MED) to encourage the recipient to pick a particular exit point for sending traffic to the neighboring AS.
- *Local preference:* An iBGP message may include a local preference attribute to aid the recipient in ranking the paths learned from different routers in the AS.
- *Community:* The community attribute provides a generic mechanism for tagging routes to aid in specifying and applying routing policies. For example, an AS might assign different community values to a path depending on whether it was learned from a customer or a peer.

These attributes play an import role in the BGP decision process, as discussed in the next subsection.

2.2 Path Selection

A BGP-speaking router may learn multiple paths to the same destination prefix from eBGP and iBGP neighbors. Although the selection of a best path depends on the attributes in the BGP update messages, the complete details of the decision process

[1] The simplest way to convey routing information throughout the backbone is to have an iBGP session between each pair of routers (i.e., a full iBGP mesh). However, the full-mesh approach introduces considerable overhead in a large backbone network. Instead, a large AS may employ techniques such as route reflectors or confederations to distribute BGP advertisements in a hierarchical fashion [8].

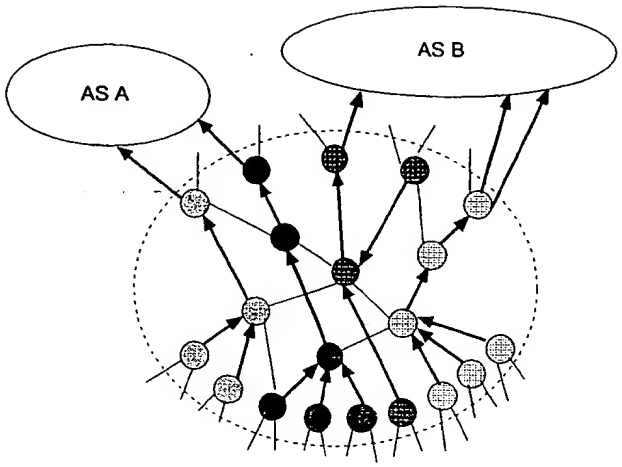

Figure 1: Flow of traffic from ingress routers to the egress links

are not part of the protocol specification. Nevertheless, router vendors adhere to a *de facto* standard [12–14] to facilitate interoperability between different products. First, certain routes are excluded from consideration. This includes the routes removed during import processing (e.g., by a route filter or due to loop detection) and routes that have an unreachable next hop. Then, the router applies a sequence of steps to narrow the set of candidate routes to a single choice, as follows:

1. *Highest local preference*: Prefer a route with the highest local preference, where local preference is assigned by the import policy and is conveyed via iBGP.
2. *Shortest AS path*: Prefer a route with the shortest AS path length, as conveyed in the BGP advertisement.
3. *Lowest origin type:* Prefer a route with the lowest origin type (IGP is preferable to EGP which is preferable to INCOMPLETE), as conveyed in the BGP advertisement or reset by the import policy.
4. *Lowest MED:* For routes with the same next-hop AS, prefer a route with the smallest MED value, as conveyed in the BGP advertisement or reset by the import policy.
5. *eBGP over iBGP*: Prefer a route learned via eBGP over routes learned via iBGP, since leaving the AS directly is preferable to forwarding traffic through the AS to another router.
6. *Lowest IGP metric*: Prefer a route with the smallest intradomain (Interior Gateway Protocol) metric to reach the next hop, since this enables each router to select its "closest" exit point.
7. *Oldest route*: Prefer the route that was received earliest, since this route is more likely to be stable.
8. *Lowest router id*: Prefer the route learned from a router with the lowest router identifier, as conveyed during establishment of the BGP session.

Most router vendors have configuration options for disabling one or more of these steps; some vendors also have support for additional steps.

The construction of the forwarding table at each router depends on the complex interaction of BGP routing policies, the distribution of update messages via iBGP, and the IGP parameters. Over time, each router receives eBGP messages from neighboring domains and iBGP messages for the best routes seen at other routers in the AS. In the meantime, the routers also participate in an IGP that affects their selection of the best path, as well as the route through the domain to reach the BGP next hop. Figure 1 shows a collection of routers that select different routes toward a destination prefix reachable via AS's A and B. Each router selects a route with the "closest" egress point, based on the IGP weights (in step 6 of the BGP decision process). Modeling the impact of interdomain routing on the flow of traffic in the network requires a way to separate the roles of BGP policies and IGP parameters in the construction of the forwarding table. It also requires a way to capture how the asynchronous exchange of eBGP and iBGP messages affects the selection of the best path at each router.

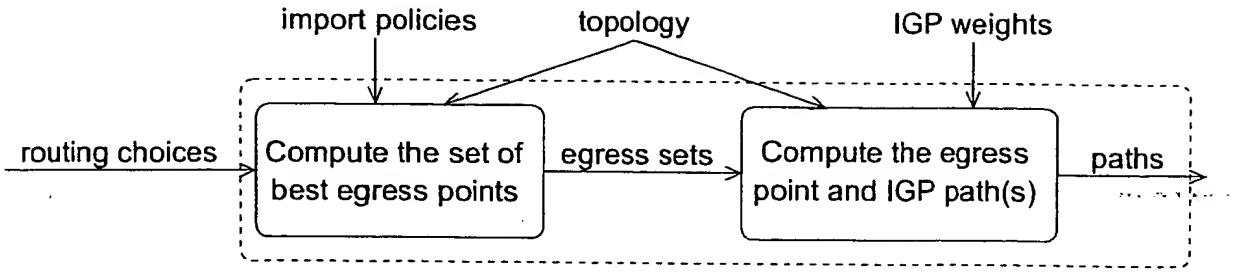

Figure 2: Modeling the impact of BGP policies and IGP weights on the flow of traffic

3 Modeling BGP Routing Choices

The selection of the best route to a destination prefix depends on the BGP advertisements, the import policies, the IGP weights, and the BGP decision process. The distributed, asynchronous nature of the routing protocols makes it difficult to predict how changes in BGP import policies affect the flow of traffic. In this section, we first present a deterministic representation of the BGP routing choices for each destination prefix. This network-wide representation accounts for the vagaries of the BGP decision process and the interplay between BGP and IGP. Next, we describe how to populate this representation from the BGP routing tables available in an operational network. Then, we outline key assumptions and principles underlying the application of the routing model in Figure 2.

3.1 Network-Wide Representation of BGP Routing Choices

Predicting the impact of configuration changes on the flow of traffic requires an accurate model of the interaction of BGP policies and IGP weights over a distributed collection of routers. The IGP parameters influence IP routing in two main ways—by affecting the BGP decision process (in step 6) and by determining the paths between the routers within the AS. The first box in Figure 2 isolates the portions of the BGP decision process that do not depend on the IGP weights. This box captures the selection of the best routes learned from neighboring domains. For each destination prefix, this produces a set of best routes (and the associated egress links), where the final selection of a best path may vary at different routers inside the AS, as shown in Figure 1. The second box captures the selection of the closest egress point, based on the IGP cost and the final tie-breaking steps (steps 6-8) for each router in the domain. This box also identifies the IGP path(s) associated with the minimum cost. This determines how traffic that enters at a particular ingress point travels to a certain egress point en route to the destination prefix. By combining this information with traffic measurements from the ingress points, the model in Figure 2 can be used to predict how a change in import policies and/or IGP weights would influence the volume of traffic on each link in the network.

Previous research on intradomain traffic engineering has shown how to compute the shortest path from an ingress point to a set of egress links based on the topology and the IGP weights [5]. This provides the basis for the second module in Figure 2. Our work focuses on the first module, with an emphasis on (i) how to represent the *routing choices* and (ii) how to ensure that changes in import policies have a *predictable* influence on the flow of traffic. As outlined in Section 2.2, the best route for a particular prefix is determined by applying import policies to a sequence of BGP updates and applying the BGP decision process to determine the best routes. In reality, the process of selecting the best route for each router is asynchronous and distributed. Rather than simulating the exchange of BGP messages throughout the network, we propose a static, centralized representation of the routing choices and the BGP decision process. This enables a network operator to predict the influence of changes in BGP import policies, based on a snapshot of the current set of advertisements received from neighboring domains. In the rest of this subsection, we propose a simple representation of the routing choices.

An effective representation of the routing choices in Figure 2 should capture the key BGP attributes that affect the flow of traffic. Router configuration languages provide considerable flexibility in specifying import policies. Some aspects of import policy, such as route filtering, do not relate directly to traffic engineering. In addition, some steps in the BGP decision process depend on BGP attributes that are largely beyond the control of the network operator. For example, AS path length (step 2) depends on the BGP advertisement as heard from a neighbor, and the eBGP/iBGP distinction (step 5) depends on whether the BGP neighbor that sends the advertisement is within the AS or not. The later stages in the decision process (steps 6–8) do not relate directly to BGP import policy. The import policy impacts the decision process primarily by setting local preference (step 1), accepting or resetting the origin type (step 3), and accepting or resetting MEDs (step 4). Local preference offers the most

$|l| = 1$: $r_1 : (p_1); r_2 : (p_2); r_3 : (p_3)$
$|l| = 2$: $r_2 : (p_4, p_5); r_3 : (p_6)$
$|l| = 3$: $r_1 : (p_7, p_8); r_2 : (p_9); r_3 : (p_{10})$

Figure 3: Representation of routing choices for a destination prefix

flexibility, since this attribute affects the first stage in the decision process and because configuration languages allow operators to assign local preference in a variety of ways. For example, network operators can assign local preference based on a regular expression match on the AS path associated with each route advertisement.

Most of the steps in the BGP decision process depend on either BGP import policy or IGP weights, except for the step based on AS path length. This step in the decision process forces all of the best paths in the set of egress points for a given prefix to have the same AS path length. Best paths of different lengths cannot coexist[2]. Hence, we *group the route advertisements based on AS path length* in order to create sets of possible best routes to a particular destination prefix. Suppose the AS learns a set of paths P_d to destination prefix d from neighboring AS's. For a path $p \in P_d$, let $|p|$ represent the AS path length. A path $p \in P_d$ of length $|p| = l$ is not selected as a best path unless all paths with shorter length have been assigned a *lower* local preference value. Within the set paths of length l, we *group advertisements by the router* $r(p)$ that received the advertisement from a neighboring AS. The BGP decision process requires each router to select a single best path, even if multiple routes have the same AS path length. If a router learns multiple routes with the same AS path length, we *order these routes by the identifier of the neighboring router*, since this determines how the router would break a tie in step 8 of the decision process.

For example, suppose an AS has three routers—r_1, r_2, and r_3—that have eBGP sessions with neighboring AS's. Suppose each router learns ten routes $p_1, p_2, \ldots, p_{10}$ *to one particular prefix* with three different AS path lengths. We group these routes into sets as shown in Figure 3. For a single prefix, each router has one path of length 1, representing a possible egress set (p_1, p_2, p_3) of best paths for this destination prefix. Depending on the local preference, MED, and origin type, the actual set of *best* routes to that prefix may be a *subset* of (p_1, p_2, p_3). For example, an import policy that assigns a low local preference to p_1 would reduce the set of best routes to (p_2, p_3). In the case where we use local preference to force a router to select its best route from routes where $|l| = 2$, router r_2 has two routes of length 2 (p_4 and p_5); ultimately, r_2 must select at most one of these two routes as its "best" route for that prefix and path length.

Effectively, the local preference assigned in the import policy determines which *row* in Figure 3 contributes routes to the egress set. Then, the local preference, origin type, MED, and router identifier determine which *entries* in this row actually appear in the egress set. Ultimately, each prefix is associated with best routes (and associated egress links) at one or more routers. This egress set represents the outcome of the first five steps of the BGP decision process at each of the routers in the network. The egress set serves as the input to the second box in Figure 2. The decision that each router makes for its best route depends on its view of the IGP costs (step 6) and, if necessary, the identifier of the router responsible for advertising the route (step 8). These later stages in the decision process can be emulated by computing the IGP path costs based on a network-wide view of the topology and the intradomain routing configuration, as captured in the second module of Figure 2. Combining the BGP and intradomain routing models, a network operator can predict how traffic for a given destination prefix would travel from a particular entry point through the network to a single egress point.

3.2 Populating the Model From BGP Routing Tables

Ideally, the network operator would have a complete, up-to-date snapshot of all of the BGP updates heard from eBGP neighbors. This would enable the operator to determine precisely how a change in import policies would affect the routing decision made by each router. However, acquiring a timely view of all of the BGP update messages in the network may be difficult in practice. Some routers can be configured to provide a continuous feed of all of the routes, both best and alternate paths, as they arrive [15], but this feature is not universally available. An alternate approach is to extract the set of paths from the BGP routing table (the Routing Information Base) from each router at the edge of the network. A simple script can telnet or ssh to each router to apply a command, such as "show ip bgp" in Cisco IOS parlance, to dump the current routing table. Figure 4 shows an example line in a BGP routing table. The entry lists a single route for prefix 38.138.55.0/24 that was learned via iBGP (the "i" before the prefix) and has a next-hop IP address of 192.168.0.10. The routing table entry includes other attributes such as the MED value (2130), local preference (100), AS path (1 701 17031), and the origin type ("i" for IGP). The ">" symbol indicates that this is the router's "best" route for this prefix.

[2]Later, in Section 4.3, we propose an effective way to relax this restriction.

```
   Network          Next Hop            Metric LocPrf Weight Path
*>i38.138.55.0/24   192.168.0.10           2130    100      0 1 701 17031 i
```

Figure 4: Example BGP routing table entry for prefix 38.138.55.0/24

Extracting the paths from routing table dumps has two main limitations regarding the quality of the data. The first limitation concerns the *accuracy* of the routing table data. Dumping the entire routing table imposes a load on the router, making it impractical to collect routing tables very frequently. In fact, since routing table dumps do not occur instantaneously, the state of the table may change during the dump itself; most router implementations avoid this problem by deferring changes in the routing table until the dump is complete. Tables collected from different routers may not represent the exact same moment in time, resulting in occasional inconsistencies in the network-wide view of the routing choices. The significance of these issues depends on how often routing changes occur relative to the frequency of the routing table dumps. Given that many routes are stable for days or weeks at a time [16], this may not be a major concern. In the long term, though, augmenting routing table data with live feeds of BGP updates would help improve the accuracy of the data.

The second limitation concerns the *completeness* of the data. The routing table represents the collection of routes *after* the import policies have been applied[3]. Hence, the table does not include any routes filtered by the import policy. Since we do not try to model changes in the filtering policy, this is not a significant limitation. Each routing table entry includes attributes such as local preference, MED, and origin type *after* manipulation by the existing import policy. This does not preclude experimenting with different import policies that change the assignment of local preference or origin type, or that reset the MED value. Finally, routing table entries such as Figure 4 do not include the communities included in the BGP advertisement. As such, the BGP tables obtained by such methods are not useful for experimenting with import policies based on communities. Instead, we focus our attention on policies that set the local preference and origin type attributes based on the prefix and the AS path. Despite these shortcomings, routing table dumps provide a view of the network that is accurate enough to derive a reasonable representation of routing choices.

We collected BGP routing table dumps from routers that connect the AT&T IP Backbone to its peers and parsed each table to extract the route for each prefix, focusing on the next-hop IP address, MED value, and AS path attributes. For each table, we focused on the routes learned via eBGP and ignored the routes that were propagated from other routers via iBGP. To focus on routes that use the peering links, we excluded prefixes that are reached directly by connections to customers of the AT&T backbone. Modifications to import policies for traffic engineering on peering links should not compromise the local preference assigned to customer routes. Suppose a prefix has routes learned from both customers and peers. If the customer route has a high local preference, then we do not include any of the routes for this prefix in our analysis, since traffic to this prefix should travel via the customer link(s) rather than peering links. On the other hand, if the customer route has a low local preference (indicative of a backup route), then we include the routes learned from peers, since traffic to this prefix should travel via the peer link(s) rather than the customer links.

3.3 Operating Guidelines

Modeling the influence of import policy on path selection addresses only part of the BGP traffic engineering problem. The network operator must select the import policies from a wide array of configuration options. BGP is a policy-based routing protocol that provides network operators with a great deal of flexibility in matching and assigning the attributes in the advertisement messages. However, this flexibility permits an operator to make ineffectual or even harmful changes in an attempt to shift traffic from one path to another. This section identifies several principles to help a network operator exert effective control over the flow of traffic. The first three issues address the predictability of changes in traffic flow as a result of adjustments to import policy:

First, *the set of BGP advertisements from neighbors should be relatively stable.* Predicting the influence of policy changes depends on knowing the set of advertisements announced by the neighboring AS's. Frequent changes in these advertisements make it difficult to apply the predictions of the model to the operational network. The instability of BGP advertisements has been a subject of concern in recent years [18, 19]. These studies identified vendor implementation decisions that contribute to the high volume of BGP update messages. This has led to changes in BGP implementations that have helped reduce the number of update messages. However, routing changes still occur for a variety of reasons, including equipment failures, reset BGP

[3] If "soft reconfiguration" is enabled [17], it is possible to dump the routes as they appear *prior* to import processing. For example, Cisco IOS includes a "show ip bgp received-routes" command for this purpose.

sessions, and configuration changes. Despite the large number of BGP update messages, most routes are stable for weeks at a time [16]. In our work, we assume that most traffic travels on routes that are stable on the timescale of hours. In fact, the model in Figure 2 can be used to evaluate the impact of changes in the BGP advertisements on the flow of traffic through the network, making the model useful even when the assumptions are violated.

Second, *the BGP decision process should be deterministic.* In our framework, we apply candidate BGP import policies to the set of advertisements received from BGP neighbors (i.e., the "routing choices"). Any dependence on the order or timing of these messages makes the selection of the best path difficult to predict. The BGP decision process outlined in Section 2.2 has two potential sources of non-determinism—in steps 4 and 7. In step 4, the comparison of MEDs applies only to paths learned from the same next-hop AS. This can make the selection of the best path dependent on the *order* of the comparison between paths, as illustrated by an example in [20]. Router vendors recommend enabling a configuration option ("bgp deterministic-med") for deterministic path selection in the presence of MEDs. In step 7, the router favors the least recently learned path, which makes the decision process depend on the arrival order of the advertisements. Disabling this step forces a deterministic tie-breaking process based on the router identifier (step 8)[4]. Our methodology assumes that each router enables deterministic MED comparison and disables step 7 of the decision process.

Third, *policy changes should not have an unpredictable impact on how and where traffic enters the AS.* Changes to import policy are intended to influence how traffic *exits* the AS. However, a change in import policy may cause one or more routers to advertise a new best path to downstream domains. Depending on the import policies of downstream neighbors, this could change their routing decisions and, in turn, alter whether and where traffic *enters* the network. This suggests that small, incremental modifications to the import policies are preferable to large, network-wide changes. For example, an operator might try to move a portion of the traffic on a congested peering link to a less-congested peering link. This reduces the impact on downstream customers and allows the operator to observe the influence on the flow of traffic before making additional policy changes. In addition, certain types of policy changes are less likely to influence the routing choices of neighboring domains. For example, moving traffic between two routes with the same AS path (but different egress links) does not change the BGP attributes of the routes advertised to other neighbors. We study this issue in more detail in Section 4.2.

Additionally, changing BGP import policies based on the model in Figure 2 requires careful consideration of potentially broad effects on the operational network. Taking this into consideration, a network operator should abide by the following principles:

First, *tuning the import policy should not introduce significant overhead.* A router applies the import policy to filter and manipulate advertisements as they arrive as part of constructing the Routing Information Base (RIB). In the worst case, applying a new import policy would require the router to *reset* the session with the BGP neighbor in order to receive a fresh copy of the advertisement messages. This would introduce substantial overhead on both routers and would cause temporary routing instability that could spread to other parts of the Internet. To avoid this problem, network operators typically configure the routers to store a local copy of each received advertisement. Enabling the "soft reconfiguration" feature on inbound routes allows the router to apply the new import policy without disrupting the BGP session to the neighbor [17]. We assume that this feature is enabled on the BGP-speaking routers in the AS. Still, applying the new import policy does incur an overhead on the router and may trigger the selection of a new best path for some prefixes; the router, in turn, may need to advertise these new paths to its other BGP neighbors. As a result, network operators should minimize the frequency of policy changes and the number of prefixes affected.

Second, *traffic engineering actions should not alter the relationship with the neighboring AS's.* Some changes to import policy fall beyond the scope of traffic engineering. For example, suppose that an AS has an agreement to accept MEDs from a neighbor. Changing the import policy to reset the MEDs, or using a filter or local preference to alter the selection of the egress point, would violate this agreement. Similarly, network operations practices or agreements with neighbors often impose constraints on the relative preference of routes learned from customers, peers, and providers. For example, network operators typically prefer routes learned from downstream customers over routes learned from peers and upstream providers [10,22]. Alternatively, a customer may request that a path be treated as a backup route by giving preference to other paths. These constraints can be obeyed by defining a distinct *range* of local preference values for each class of routes and by requiring the import policies to adhere to these restrictions. Finally, the relationship with a neighboring AS may impose restrictions on the volume of traffic exchanged in each directions. Certain changes in routing policy may violate these agreements. In fact, our model can be used to detect when a proposed change in import policy might lead to violations of traffic agreements with neighboring domains.

[4] Other BGP features, such as route flap damping [21], can help avoid repeated advertisement and selection of unstable paths.

4 Guidelines for BGP Traffic Engineering

Our routing model in Figure 2 predicts how changes in import policies affect how traffic flows through the network. However, the number of possible policy changes is extremely large, and the BGP decision process imposes limitations on how a network operator can control the flow of traffic. In this section, we analyze routing and traffic data to identify practical approaches for tuning import policies to the prevailing traffic patterns. Specifically, we:

- consider ways to *scope the problem* by focusing on a small number of prefixes, or groups of related prefixes,
- *evaluate various techniques for shifting traffic* between different egress links, either to the same neighboring AS or between links to different neighbors,
- show how a network operator can redirect traffic while *limiting the influence on the routing decisions in neighboring domains*, and
- propose ways to incorporate AS path length into the BGP decision process *without requiring all best paths to have the same length.*

Our analysis draws on BGP tables from the routers that connect the AT&T IP Backbone to other large providers. We used the routing tables to construct the set of *routing choices* for each destination prefix reachable via one or more peering links, as discussed in Section 3.2 and shown in Figure 3. Then, we analyzed the characteristics of these routing choices and the implications on how network operators can change BGP import policies to move traffic from one location to another. Because traffic volume is not evenly distributed across destination prefixes, we also analyzed routing choices with respect to outbound traffic volume, paying attention to how much traffic is destined to each destination prefix and AS. We analyzed the outbound traffic volume from a subset of these routers using Cisco's Netflow feature [23], aggregating these statistics for each destination prefix over a day[5]. We collected BGP routing tables at approximately 2 a.m. EDT on June 1, 2001 and the Netflow measurements throughout the day on June 1, 2001. Additionally, we collected the same data on June 21, 2001 to verify the results of the analysis on data from a different day. The AT&T IP Backbone is a large AS with no upstream providers. As such, the specific traffic statistics may differ for other types of networks or for different parts of the AT&T network. Nevertheless, we expect the basic trends and general principles we observe to apply to other AS's.

4.1 Limiting the Scale of the Problem

Because a typical default-free routing table contains routes for more than 90,000 prefixes, exploring all possible combinations of BGP import policies is computationally intractable. However, by focusing on the small fraction of "popular" prefixes, or alternatively on sets of prefixes that can be grouped according to common BGP attributes, a network operator gains significant flexibility for traffic engineering, while avoiding overly complex import policies and an incredibly large number of routing policy options.

4.1.1 Group Prefixes with the Same Routing Choices

Import policies that are tailored to every prefix at every router would be extremely complicated to configure and expensive for the router to apply. In addition, such fine-tuned policies might not remain appropriate following a shift in traffic or a change in the neighbors' routing updates. Fortunately, many prefixes have the same attributes across all eBGP advertisements from neighboring domains. With regard to the representation in Figure 3, this corresponds to grouping the prefixes that have the same representation for routing choices. The idea of grouping prefixes with the same routing choices was also proposed in [24]; however, this previous work considers the BGP advertisements in a single routing table, rather than constructing a *network-wide view* of the routing choices within an AS across multiple routers.

For the routers connecting the AT&T IP Backbone to its peers, we find a total of 27,000 unique representations of routing choices. On average, a set of routing choices is associated with three destination prefixes. However, the number of related prefixes is much larger in some cases; in one case, 1048 destination prefixes had exactly the same routing choices. These results

[5]Each Netflow record includes the destination IP address and the mask length for the longest-matching prefix in the routing table. We associate the volume of traffic in each record with the corresponding destination prefix.

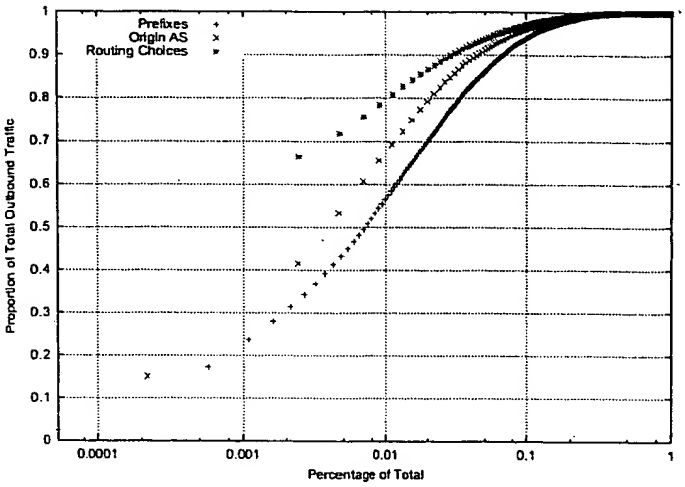

Figure 5: Distribution of traffic over outbound peering links on a peering router for the top destination prefixes over a 24-hour period. For the set of routers on which we collected Netflow data, 10% of the destination prefixes are responsible for 95% of the outbound traffic to peering links.

suggest that one way to reduce the scale of the problem is to choose egress links for traffic based on unique routing choices, rather than by destination prefix. Given a group of prefixes with the same routing choices, a network operator can manipulate all of these prefixes *together* by assigning attributes (such as local preference) to these advertisements based on their common attributes, such as AS path characteristics. Additionally, studies have shown that aggregated traffic may be more stable than the traffic associated with any particular destination prefix [25] . Thus, by specifying policy by grouping advertisements with common attributes, a network operator can shift a group of prefixes, rather than having to adjust policies on a prefix-by-prefix basis.

4.1.2 Focus on Popular Prefixes

Defining independent import policies even for 27,000 unique routing choices is still an unreasonable requirement. Fortunately, a large fraction of the traffic is concentrated in a small fraction of the prefixes. The bottom curve in Figure 5 shows the cumulative distribution of the proportion of traffic contributed by the most popular prefixes. For example, traffic destined for the top 0.1% of the prefixes is responsible for more than 20% of the outbound traffic. The top 10% of prefixes accounts for approximately 95% of the traffic. Similar results have been seen in other traffic measurement studies [25–27] . The results are even more dramatic when we *group prefixes with the same routing choices*, as shown by the top curve in Figure 5. For example, 1% of the 27,000 sets of routing choices contribute more than 70% of the traffic. Even grouping traffic by common origin AS results in a significant concentration of traffic—1% of origin AS's are responsible for 70% of outbound traffic to peers.

These results suggest that a network operator can exert significant control over the flow of traffic through an AS by focusing on a relatively small number of popular prefixes or sets of prefixes. Changing import policies for a small group of prefixes also limits the number of new routing advertisements sent to neighboring domains. In this case, a network operator selects *specific* prefixes (i.e., from those which carry large portion of the traffic), or a specific group of related prefixes.

4.2 Shifting Traffic Away from a Congested Link

A network operator can alleviate congestion on an edge link by directing a portion of the traffic to another link, either to the same neighbor AS or to a different AS. Shifting traffic between links to the same AS prevents changes in the volume of traffic (and the ratio between inbound and outbound traffic) sent to each neighboring domain and reduces the likelihood that downstream AS's receive new route advertisements that would inadvertently change how traffic enters the network. In this subsection, we identify how to shift a subset of the outbound traffic without changing the next-hop AS that carries the traffic. We then describe how the appropriate import policy for redirecting traffic depends on the characteristics of the AS paths advertised by the neighboring

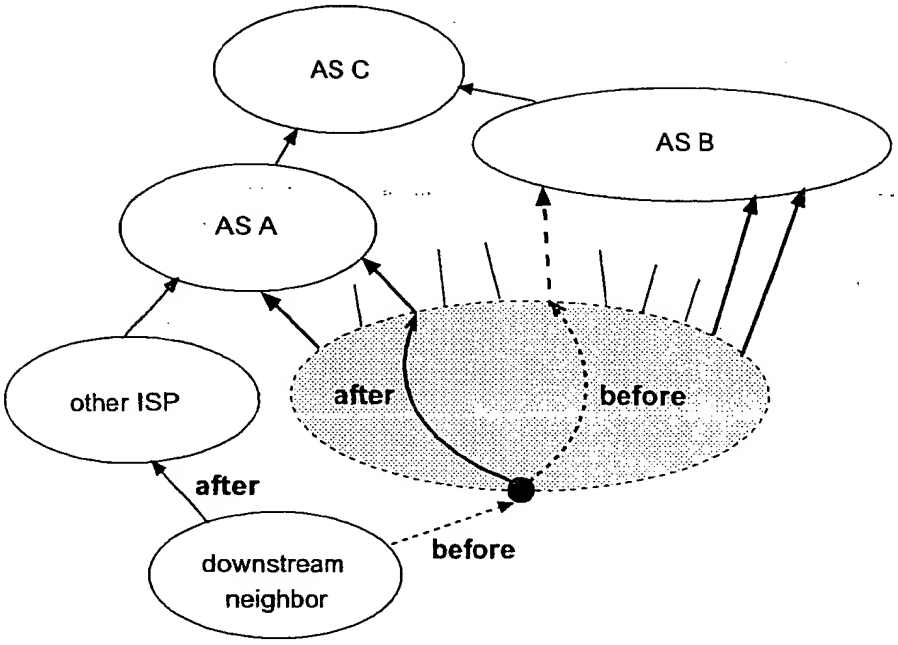

Figure 6: Change in neighbor's behavior upon receiving a new route

domain.

4.2.1 Limiting the Impact on Downstream Neighbors

In adapting routing policies, operators should minimize the potential impact on the behavior of downstream neighbors. If neighboring AS's change their behavior (i.e., routing advertisements or best route decisions) in response to a policy change in our network, the influence of a policy change on the flow of traffic through our network will be unpredictable. For example, suppose that a particular edge link is congested and the network operator assigns a lower local preference value to some of the routes traversing the congested link. The new import policy will remove the link from the egress set for one or more prefixes, thus causing some routers in the network to direct traffic for these destination prefixes to a different link in the egress set. Moving the traffic reduces the load on that congested link. However, the affected routers might advertise a new route to their eBGP neighbors, such as downstream customers, potentially causing significant changes in the inbound traffic seen by the network that made the policy change.

For example, one of the edge routers in Figure 6 might switch from a route via AS B to a route via AS A. Suppose AS's A and B advertise a path to a destination prefix in AS C. Then, the network would receive route (A, C) on the west coast and route (B, C) on the east coast. Decreasing the local preference of one of the (B, C) routes (as shown by the dashed line) would cause some routers to redirect traffic to an (A, C) route—the new "closest" egress point based on the IGP tie-break in stage 6 of the BGP decision process. These routers would advertise the new best path to downstream neighbors. Depending on the neighbor's routing policies, the new advertisement might cause the neighbor to select a different next-hop AS (e.g., another ISP) for reaching this destination prefix. This could result in a sudden and unpredictable decrease in the volume of traffic entering the domain at this router. Similarly, the routing change could trigger an increase in traffic if other neighbors preferred the (A, C) route over the (B, C) route.

To prevent routing changes in neighboring domains, the network operator should focus on sets of prefixes for which every "best" route for that prefix has the same BGP attributes (except for the next-hop IP address). Formally, from Figure 3 this means performing adjustments on prefixes where the entire row of "best" paths have the same characteristics. Depending on the BGP implementation, the downstream AS's may not even have to receive a new BGP advertisement, since none of the attributes have changed. For the AT&T peering links, 83.5% of the prefixes have shortest AS paths with a single next-hop AS, as shown in Figure 7; these destination prefixes represent over 45% of the outbound traffic. For these prefixes, reducing the local preference

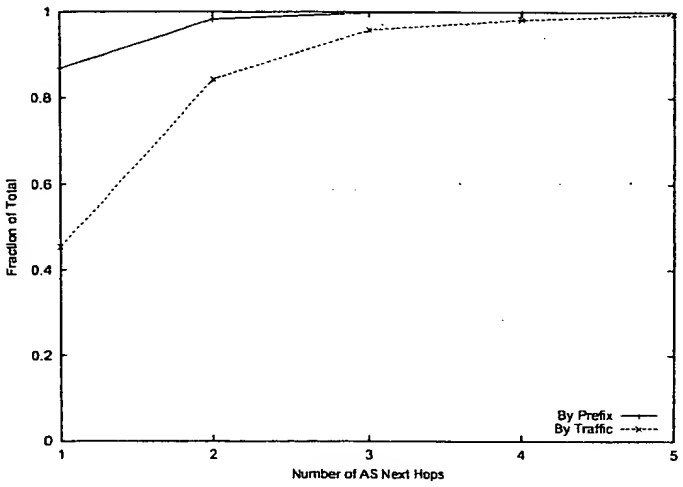

Figure 7: Cumulative distribution function of number of next-hop AS's for the shortest AS paths for a prefix. The majority of prefixes (carrying about 45% of all outbound traffic) have best routes with a single next-hop AS.

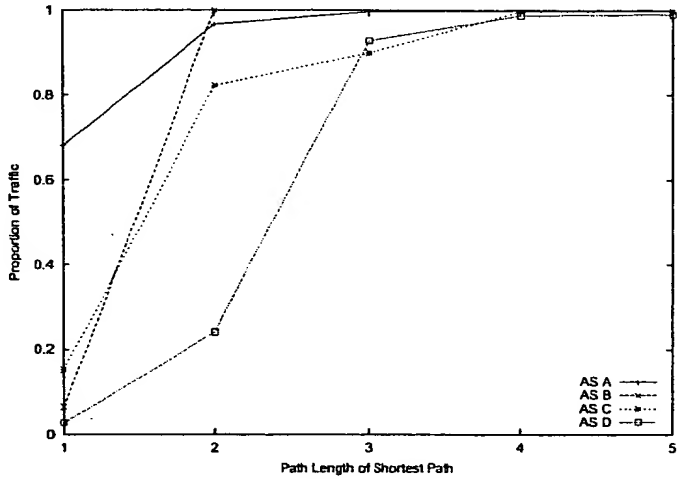

Figure 8: Cumulative fraction of outbound traffic vs. AS path length for various next hop AS's. One technique used to shed load is to assign a lower local preference to all traffic to a certain AS with a certain number of hops. This can have vastly different effects depending on which AS's advertisements are affected.

at one peering location would shift traffic to another egress link to the same peer. In some cases, a network operator may need to move traffic from one neighbor AS to another. As shown in Figure 7, a reasonable amount of prefixes and traffic have shortest paths with two neighbor AS's. This is useful for moving traffic between two neighboring AS's without having to select routes with different AS path lengths[6]. Although this type of routing change requires sending a new route advertisement to some downstream AS's, advertising a route with the *same AS path length* reduces the likelihood that a downstream AS selects a different best path.

[6] In many cases, the network may have routes to two AS's via the *same egress router*—for example, a single router may peer with both AS A and AS B. In this case, it is possible to move traffic from one egress link to another without changing the flow of traffic within the AS to reach the egress router.

4.2.2 Shifting Traffic With Simple Changes in Import Policy

Router configuration languages provide significant flexibility in assigning local preference values to routes; for example, these configuration languages allow an operator to assign a smaller local preference to routes based on the prefix or the regular expressions on the AS path. In some cases, a network operator may need to move a relatively large amount of traffic. For example, suppose that one or more links are upgraded to higher capacity. Decreasing the local preference for the low-bandwidth egress point can shift traffic to the new, high-bandwidth link more capable of carrying the traffic. Assigning local preference based on the *length* of the AS path is an effective way to move a large amount of traffic from one location to another with a small change in the import policy. This approach amounts to selecting a subset of the routing choices as in Figure 3 for a *particular value of l*. This approach is simple and does not depend on the exact sequence of AS's in the path. For example, a network operator can define an import policy for one peering session with a neighboring AS that assigns a lower local preference for routes with an AS path length of two. This would shift traffic for destination prefixes with a two-hop AS path through that neighbor to another egress point. However, the specific effects of this technique depends on how traffic is distributed over different lengths of AS paths. This may vary across different next-hop AS's.

Figure 8 shows the cumulative distribution of outbound traffic *to a particular AS* that is carried by best paths of various lengths. Each curve corresponds to a different next-hop AS, identified by A, B, C, and D; for example, nearly 70% of the outbound traffic to AS A travels over a one-hop AS path (where AS A is the next-hop AS). In contrast, the majority of traffic traveling via the other three AS's travels on AS paths of length two or three. These significant differences stem from the various roles AS's in the Internet can play, as well as historical and network-specific artifacts (e.g., often a single network will have multiple AS's). In some cases, an AS hosts a large number of services and directly-connected customers that do not have their own AS numbers. This type of network sends traffic over paths with a single AS hop, as shown in the plot for AS A. In other cases, an AS is a transit provider for a large number of tier-2 providers or multi-homed institutions. Outbound traffic to these types of networks is likely to travel over paths of different lengths, as shown in the plots for AS's B, C, and D. Depending on the diversity of next-hop AS's, a network operator should expect to see differences in the distribution of traffic over AS path lengths, which should play a role in the selection of import policies for shifting traffic to different egress links for each AS.

4.3 Controlling the Influence of BGP Attributes in Advertisements from Neighbors

The BGP advertisements from neighboring domains have a significant influence on the selection of the best paths for each destination prefix. Although the import policy is capable of resetting some attributes (such as MED and origin type), other attributes such as the AS path depend on the policies applied in other domains and cannot be reassigned by import policy. Inconsistencies in the routes advertised via different eBGP sessions with the same next-hop AS can diminish a network operator's control over the flow of traffic. In addition, the common practice of AS prepending limits a network's ability to spread traffic over a large number of egress points. In this subsection, we quantify the influence of routing policies in other domains on a network's flexibility in selecting best routes. Based on these results, we suggest techniques for increasing a network's control over the flow of outbound traffic.

4.3.1 Ensuring Consistent Advertisements from Neighbor AS's

BGP update messages from neighboring AS's have a significant impact on the flow of traffic through a network. A neighbor AS has the ability to exert influence on how traffic leaves a network by sending inconsistent routing advertisements over different eBGP sessions. For example, suppose that a network connects to AS A at locations on the east and west coast. If AS A advertises a prefix only on the east coast, then this would force the network to carry all of the outbound traffic for this prefix to the west coast. Alternatively, AS A might advertise the path with a different AS path length or origin type at different locations. Inconsistent advertisements can have a significant and unpredictable influence on the flow of traffic by limiting the number of egress points. We analyzed the routes in the BGP tables to identify paths of different lengths from the same next-hop AS for the same destination prefix. Across all prefixes and next-hop AS's, we found inconsistent path lengths in just 0.03% of (prefix, next hop AS) tuples; in addition, for 0.09% of these cases, some of the peering sessions to the next-hop AS did not advertise a prefix that was advertised at other locations. The very small number of inconsistent routes is likely due to the asynchrony in downloading the BGP tables from the routers. Although inconsistent advertisements were not significant in our data set, a network operator should still make periodic checks for consistency to ensure maximal flexibility for making routing choices.

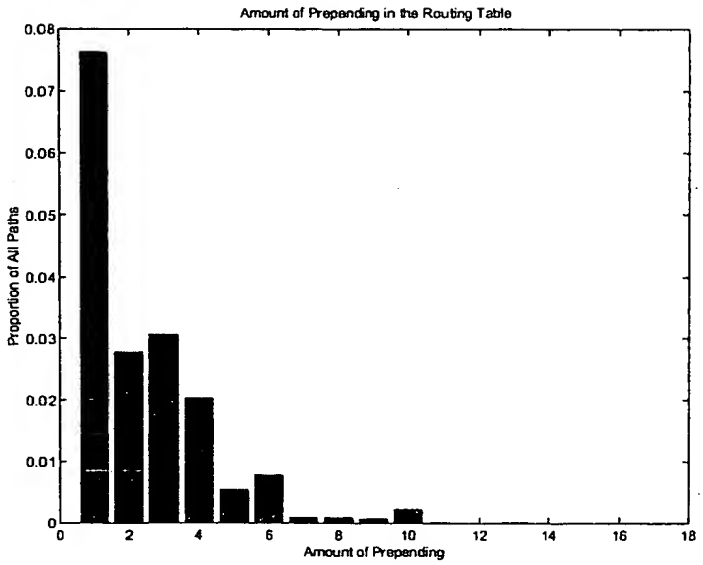

Figure 9: Number of occurrences that an AS path length was extended by the given amount by prepending. 17.4% of all advertised paths included some prepending. The majority of prepended paths were extended by one or two hops, many paths were extended by 10 hops, and 5 paths were extended by as many as 17 hops.

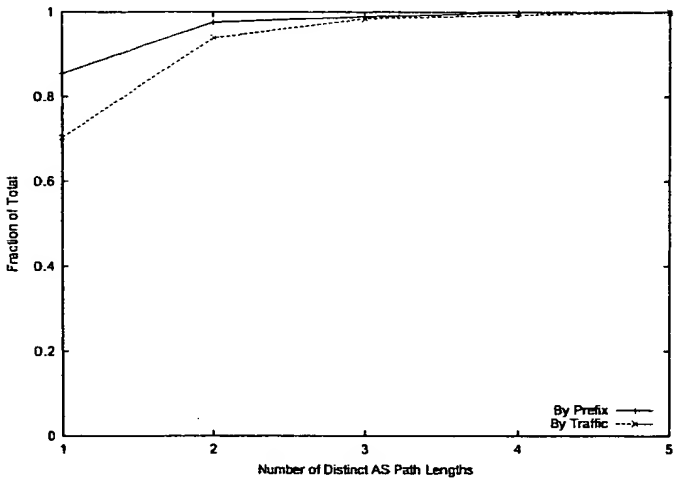

Figure 10: Distribution of prefixes for number of distinct AS path lengths. The majority of prefixes are reachable through paths of only one length, but the remaining 15% of prefixes have paths of multiple lengths. About 11% of these prefixes have two distinct path lengths, which suggests a multihomed prefix with one prepended path to specify a backup route.

4.3.2 Limiting the Influence of AS Path Length

Even if advertisements from neighbors are consistent across eBGP sessions to the same next-hop AS, AS path length has a considerable influence on the comparison of routes from different next-hop AS's. AS prepending inflates the length of the AS path by repeating an AS number multiple times to artificially make a path look longer. For example, consider an AS C that connects to providers A and B. AS C may send a one-hop route (C) to AS A and a three-hop route (C, C, C) to AS B to encourage traffic destined to AS C to traverse a route via AS A. Figure 9 shows that AS path prepending is a quite common practice. Approximately 18% of the routes had some amount of AS prepending. The majority of these paths were extended by one or two hops. AS prepending contributes to the diversity of AS path lengths, as shown in Figure 10. Over 10% of the

destination prefixes and traffic have AS paths with more than one length (on other days, this number was as high as 30% of prefixes). The different lengths stem from a mixture of AS prepending and routes with a different number of AS's in the path. In either case, the different lengths limit flexibility in selecting a set of best routes, since step 2 in the BGP decision process forces all paths in the egress set to have the same length.

Small differences in AS path length do not necessarily have a significant influence on end-to-end performance. In some cases, one AS hop may represent a large number of router hops or high propagation delay; in other cases, an AS hop may represent a single router hop or a small propagation delay. Forcing all best paths to have the same AS path length may be unnecessarily restrictive. Figure 8 shows that the majority of traffic travels over shortest AS paths of length 2 or 3. Furthermore, almost no traffic traverses AS paths of length 4 or longer. Consequently, it may be effective to allow the set of best paths to include paths of small differences in length (e.g., having one 2-hop path and one 3-hop path to a prefix may provide more flexibility than allowing only the 2-hop path). Coarse-grained AS path length categorization can be achieved by *disabling* step 2 of the BGP decision process and instead assigning local preference ranges based *in part* on AS path length. For example, a network operator could assign a range of local preference values to one-hop paths, another range to paths of length 2 or 3, and so on. This ensures that AS path length has an influence on the decision process without imposing the strict requirement that all best paths for a destination prefix must have the same length.

5 Conclusion

BGP is a flexible interdomain routing protocol that scales to the large number of AS's in today's Internet. However, BGP was not designed with traffic engineering in mind. The attributes available in BGP advertisements, the restrictions in the BGP decision process, and the constraints imposed by configuration languages all limit an operator's ability to tune routing policies to the prevailing traffic patterns. Despite these limitations, it is possible to control the flow of traffic by adhering to certain guidelines and employing a model for predicting the influence of changes in routing policies. We have proposed a concise, network-wide representation of the routing advertisements from neighboring domains and described how to populate these models from the BGP routing tables available in operational networks. Drawing on routing and traffic data from the AT&T IP Backbone, we have proposed and evaluated techniques for limiting the scope of BGP policy changes and reducing the impact of these changes on the routing decisions made in neighboring domains. We also propose a way for AS path length to influence the BGP decision process without requiring all best routes to have the same length.

By presenting a framework for interdomain traffic engineering, as well as sensible ways for controlling the impact of routing policy changes, we have opened a variety of avenues for future work:

- **Handling inbound traffic:** In this paper, we have focused on the influence of BGP import policies on *outbound* traffic; however, a complete solution should consider inbound traffic as well. Since an operator has limited control over how traffic enters the network (using crude techniques such as AS prepending), we believe that neighboring AS's should coordinate to gain a greater level of predictability with respect to how traffic enters each network. We are considering ways for neighboring AS's to cooperate via inband signaling (e.g., using the BGP community attribute) without revealing their network topologies and routing policies.

- **Defining the performance objective:** Traffic engineering involves tuning routing policies based on a target performance objective. The commercial relationships between AS's impose constraints and costs based on the volume of traffic exchanged with neighboring domains. In addition, the distribution of traffic after network failures may also play a role in evaluating possible changes to the routing configuration. Drawing on earlier work on IGP optimization, our ongoing work considers new objective functions that capture the constraints of both intradomain and interdomain routing, including the influence of peering agreements.

- **Incorporating end-to-end performance:** Changes in BGP routing policy affect the *end-to-end* path from a source to a destination which, in turn, influences communication performance. We are investigating ways to collect information about the performance properties of the rest of the path to help weigh the benefits of different changes in BGP policies and IGP weights. For example, active measurements that identify congestion problems in other AS's would lend insight into which policy changes would improve end-to-end performance.

These ongoing research efforts can draw on the representation of routing choices and the insights from the analysis of routing and traffic measurements presented in this paper.

Acknowledgments

We would like to thank Tim Griffin for many very helpful discussions. Thanks also to Dave Andersen, Hari Balakrishnan, Randy Bush, Steve Garland, Joel Gottlieb, Carsten Lund, and Aman Shaikh for helpful feedback on earlier versions of the paper.

References

[1] V. Paxson, "End-to-End Routing Behavior in the Internet," *IEEE/ACM Trans. Networking*, vol. 5, pp. 601–615, October 1997.

[2] D. O. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao, "Overview and principles of Internet traffic engineering." Work in progress, Internet-Draft draft-ietf-tewg-principles-01.txt, October 2001.

[3] D. O. Awduche, J. Malcolm, J. Agogbua, M. O'Dell, and J. McManus, "Requirements for traffic engineering over MPLS." Request for Comments 2702, September 1999.

[4] D. O. Awduche, "MPLS and traffic engineering in IP networks," *IEEE Communication Magazine*, pp. 42–47, December 1999.

[5] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, and J. Rexford, "NetScope: Traffic engineering for IP networks," *IEEE Network Magazine*, pp. 11–19, March 2000.

[6] X. Xiao, A. Hannan, B. Bailey, and L. Ni, "Traffic engineering with MPLS in the Internet," *IEEE Network Magazine*, pp. 28–33, March 2000.

[7] Y. Rekhter and T. Li, "A Border Gateway Protocol." Request for Comments 1771, March 1995.

[8] S. Halabi and D. McPherson, *Internet Routing Architectures*. Cisco Press, second ed., 2001.

[9] J. W. Stewart, *BGP4: Inter-Domain Routing in the Internet*. Addison-Wesley, 1998.

[10] G. Huston, "Interconnection, peering, and settlements," in *Proc. INET*, June 1999.

[11] B. Fortz and M. Thorup, "Internet traffic engineering by optimizing OSPF weights," in *Proc. IEEE INFOCOM*, March 2000.

[12] "BGP Best Path Selection Algorithm." `http://www.cisco.com/warp/public/459/25.shtml`.

[13] "How the Active Route Is Determined." `http://arachne3.juniper.net/techpubs/software/junos42/swconfig-routing42/html/protocols-overview4.html#1045417`.

[14] "Foundry Switch and Router Installation and Configuration Guide, Chapter 19, Configuring BGP4." `http://www.foundrynet.com/services/documentation/SRguide/FoundryManual_BGP4.html`.

[15] "Have BGP Advertise Nonactive Routes." `http://www.juniper.net/techpubs/software/junos42/swconfig-routing42/html/bgp-config38.html#1015169`.

[16] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental study of Internet stability and wide-area network failures," in *Proc. International Symposium on Fault-Tolerant Computing*, June 1999.

[17] "BGP Soft Reset Enhancement." `http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newft/120t/120t7/sftrst.htm`.

[18] C. Labovitz, R. Malan, and F. Jahanian, "Internet routing stability," *IEEE/ACM Trans. Networking*, vol. 6, pp. 515–528, October 1998.

[19] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet routing convergence," *IEEE/ACM Trans. Networking*, vol. 9, pp. 293–306, June 2001.

[20] "How BGP Routers Use the Multi-Exit Discriminator for Best Path Selection." `http://www.cisco.com/warp/public/459/37.html`.

[21] C. Villamizar, R. Chandra, and R. Govindan, "BGP Route Flap Damping." Request for Comments 2439, November 1998.

[22] L. Gao and J. Rexford, "Stable Internet routing without global coordination," in *Proc. ACM SIGMETRICS*, June 2000.

[23] Cisco Netflow. `http://www.cisco.com/warp/public/732/netflow/index.html`.

[24] A. Broido and K. Claffy, "Analysis of RouteViews BGP data: Policy atoms," in *Workshop on Network-Related Data Management*, May 2001.

[25] N. Taft, S. Bhattacharyya, J. Jetcheva, and C. Diot, "Understanding traffic dynamics at a backbone POP," in *Proc. Workshop on Scalability and Traffic Control in IP Networks, SPIE ITCOM+OPTICOMM Conference*, August 2001.

[26] W. Fang and L. Peterson, "Inter-AS traffic patterns and their implications," in *Proc. IEEE Global Internet Symposium*, December 1999.

[27] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, "Deriving traffic demands for operational IP networks: Methodology and experience," *IEEE/ACM Trans. Networking*, vol. 9, June 2001.

AT&T

Document Cover Sheet for Technical Memorandum

itle: Controlling the Impact of BGP Policy Changes on IP Traffic

Authors	Electronic Address	Location	Phone	Company (if other than AT&T)
Nick Feamster	feamster@lcs.mit.edu			MIT
Jay Borkenhagen	jayb@att.com	MT C5-3D12	732-420-2526	
Jennifer Rexford	jrex@research.att.com	FP A169	973-360-8728	

Document No.
HA173000-011106-02TM

Work Project No.

eywords:
Border Gateway Protocol (BGP), traffic engineering, traffic measurement, routing policy, network operations

ERCURY Announcement Bulletin Sections
CMP- Computing

bstract

The Internet consists of nearly 12,000 autonomous systems (AS's) that exchange routing information using the Border Gateway Protocol (BGP). The operators of each network need to have control over the flow of traffic through the AS. However, BGP does not facilitate common traffic engineering tasks, such as balancing load across multiple links to a neighboring AS or directing traffic to a different neighbor. Solving these problems is difficult because the number of possible changes to routing policies is too large to exhaustively test all possibilities, some changes in routing policy can have an unpredictable effect on the flow of traffic, and the BGP decision process implemented by router vendors limits an operator's control over path selection. In this paper, we demonstrate that it is possible to *predictably* model the changes in traffic flows in response to BGP policy changes, *given that policies are adapted in a certain fashion*. Based on analysis of routing tables and traffic measurements from the AT&T backbone, we show that operators can control the scale of the traffic engineering problem by focusing on the small fraction of destination prefixes (and sets of related prefixes) responsible for the majority of traffic. Furthermore, they can make the effects of their changes more predictable by following specific policy guidelines and selecting configuration options that make the BGP decision process deterministic. This allows an operator to gain more control over network traffic *within the existing BGP framework*.

ges of Text 0 Other Pages 20 Total 20
. Figs. 10 No. Tables 0 No. Refs. 27

Mailing Label

Complete Copy **Cover Sheet Only**

uture AT&T Distribution by ITDS Release to any AT&T employee (excluding contract employees)

uthor Signatures

Nick Feamster	Jay Borkenhagen	Jennifer Rexford

r Use by Recipient of Cover Sheet:

mputing network users may order copies via `http://attlis.att.com/services/proprietary.html`.